Journal of Biomedical Informatics 42 (2009) 692–701

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Translating medical terminologies through word alignment in parallel text corpora

Louise Deléger [a,b,*], Magnus Merkel [c], Pierre Zweigenbaum [d]

[a] INSERM, UMR_S 872, Eq. 20, Centre des cordeliers, Paris F-75006, France
[b] Université Pierre et Marie Curie, Paris F-75006, France; Université Paris-Descartes, Paris F-75006, France
[c] Department of Computer and Information Science, Linköping University, Sweden
[d] CNRS UPR3251, LIMSI, Orsay F-91403, France

ARTICLE INFO

*Article history:*
Received 4 October 2008
Available online 9 March 2009

*Keywords:*
Natural Language Processing
Medical terminology
Multilinguality
Parallel corpora
Word alignment

ABSTRACT

Developing international multilingual terminologies is a time-consuming process. We present a methodology which aims to ease this process by automatically acquiring new translations of medical terms based on word alignment in parallel text corpora, and test it on English and French. After collecting a parallel, English–French corpus, we detected French translations of English terms from three terminologies—MeSH, SNOMED CT and the MedlinePlus Health Topics. We obtained respectively for each terminology 74.8%, 77.8% and 76.3% of linguistically correct new translations. A sample of the MeSH translations was submitted to expert review and 61.5% were deemed desirable additions to the French MeSH. In conclusion, we successfully obtained good quality new translations, which underlines the suitability of using alignment in text corpora to help translating terminologies. Our method may be applied to different European languages and provides a methodological framework that may be used with different processing tools.



## 1. Introduction

Medical terminologies are a necessary resource to any kind of health care information task (e.g. coding, free text indexing, information retrieval). They may be local terminologies or standardized terminologies (such as ICD-10 or SNOMED CT). The problem with local terminologies is that they are not interoperable and can only be used within a certain framework. Today, the development of health systems requires standardization and interoperability of the terminologies, on the national level but also and increasingly on the international level. In France for instance, the adoption of SNOMED International is a much debated topic.

Internationalization of medical terminologies calls for translations of medical standards. Such standards are usually available in English but not always in other languages (or not in their full versions). The common policy is to rely on the work of translators. However this manual approach is time-consuming and requires skilled specialized translators.

We propose a methodology to automatically acquire translations of medical terms based on a Natural Language Processing technique—word alignment in parallel corpora. The main idea behind this work is to make use of existing translated texts from which translations at the term level can be extracted. Indeed multilingual standard terminologies may be scarce but plenty of multilingual texts can be found as regards a specific domain.

We focus on the quality of the results retrieved through our methodology rather than on the quantity. That is to say, we chose to concentrate on acquiring a high proportion of correct translations (i.e. translations with a high precision) rather than on retrieving the highest possible number of translations. A high precision means less reviewing time for experts which is our main objective. Increasing the coverage constitutes the future next step of this work.

## 2. Background

Previous work has addressed the need of developing and extending multilingual controlled medical vocabularies. Some methods focused on morphological information to derive word translations of medical terms: Lovis et al. [1] built multilingual dictionaries based on the ICD-10 classification using morphological decomposition; those dictionaries were aimed at patient encoding systems. Marko et al. [2] mapped monolingual medical lexicons in order to create a multilingual dictionary. Claveau and Zweigenbaum [3] generated translations of morphologically related medical terms by inferring transducers. Other methods relied on already parallel medical vocabularies on which they performed word alignment: Baud et al. [4] aligned the words of the English–French ICD-10 and Nyström et al. [5,6] used various parallel terminologies to build an English–Swedish dictionary. Finally, methods

* Corresponding author. Fax: +33 (1) 53 10 92 01.
*E-mail addresses:* louise.deleger@spim.jussieu.fr, louisedeleger@hotmail.com (L. Deléger).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.03.002

relying on text corpora were also developed for translating medical terms. Chiao and Zweigenbaum [7] looked for French translations of medical terms in comparable corpora, i.e. text corpora addressing the same general topic in two different languages. Widdows et al. [8] designed a statistical vector model to match English UMLS terms with their German translations in a corpus aligned at the document level. Ozdowska et al. [9] aligned the words of a parallel English–French corpus to find French translations of MeSH terms. While the previous methods focused on single-word translation, this approach also handled the translation of certain types of multi-word expressions by using compositional translation. Word alignment in both parallel and comparable corpora was performed to extend the German version of the MeSH and help crosslingual information retrieval [10].

In this paper, we align words of a parallel corpus in order to find French translations of English terms belonging to medical terminologies. The idea is that given a medical term in English, we find its occurrence in a text corpus, then we look for its corresponding occurrence (i.e. translation) in the French part of the corpus, which gives us a candidate French translation for the initial English term. Finding the corresponding occurrence is called alignment which is the process of matching linguistic units—paragraphs, sentences, phrases or words—of two (or more) languages that are in a translation relation. The starting point is a multilingual corpus, usually parallel (texts and their exact translations) but sometimes comparable (texts with similar content). In this work, we use a bilingual parallel corpus and align it at both sentence and word level, the former being needed for accomplishing the latter, which is our real aim.

Sentence alignment looks for sentences that are translations of each other and constitutes the typical first step towards word alignment. Although it is most common that one sentence in a source language corresponds to one sentence in a target language, there are instances where one sentence is translated with two—or sometimes even three or more—sentences, and where sentences are omitted. So before working at the word level, correspondences between sentences need to be established. Word alignment, on the other hand, matches corresponding words from aligned sentences and is a far more challenging task. There is no true word-to-word correspondence and a number of issues arise which complexify the process. A word is often translated with several words (e.g. complex words and locutions), or can be omitted in the translation (e.g. grammatical words specific to a language). Idiomatic expressions are especially hard to align since they are specific to a language and usually translate very differently. Parallel sentences, though being translations of each other, can differ considerably in terms of structure and wording. In that case even a human may have trouble determining which words should be paired together. Fig. 1 shows word alignments of two English and French sentences. This case is rather straightforward and there is no major problem in doing it manually; however we can see that a single adjective (“ *lifelong*”) is translated with a whole phrase (“ *qui dure toute la vie*”), which might cause trouble for an automated approach.

## 3. Research questions

In this paper, we aim at detecting new translations of terms from medical terminologies through word alignment in parallel corpora. In doing so a number of questions naturally arise, which we try to address. The core issue being, of course, whether such a technique as word alignment is suited for our purpose of acquiring new translations of medical terms, in terms of quality, time, and quantity. To answer that question we examine the following issues:

- We investigate the quality of acquired translations. This is a twofold issue since translations have to be linguistically valid, but also medically accurate. A translation may be perfectly correct in its linguistic context but not desirable as part of a medical terminology. Evaluation of the translations must therefore look at those two points.
- Since we aim at automating the process of translation, we have to look at the amount of human intervention required by our method and whether it can be reduced.
- Finally, we look at the proportion of translated medical terms that can be extracted. This will give an indication as to where we stand in terms of coverage. Quantity is not however the main focus of our current work. We first wish to ensure that our method can select suitable translations and thus reduce the amount of manual work necessary to review the results. Further work will then aim at maximizing the coverage.

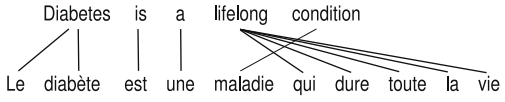


**Fig. 1.** Word alignment in two parallel sentences (taken from our corpus).

## 4. Material and methods

### 4.1. Medical terminologies

We based our work on three medical terminologies that needed to be partially or entirely translated into French: MeSH, SNOMED CT and the MedlinePlus Health Topics and that were included in the UMLS Metathesaurus. The current status of translation into French is different for each: (1) the MeSH terminology has the most advanced state of translation—each descriptor has a French translation; so further translation is only required to obtain more French synonyms; (2) SNOMED CT is only partially translated—outside of the UMLS—and a significant number of concepts still need to be translated; (3) there exists no French translation of the MedlinePlus Health Topics; it has to be translated in its integrity.

The English versions of the medical terminologies were those of the UMLS 2006 version used by the MetaMap program version 2.4.C (which we used to extract English terms from the text corpus). The French MeSH was extracted from the UMLS 2007AA. The partial French translation of SNOMED International (v3.5) was obtained courtesy of the Secrétariat Francophone International de Nomenclature Médicale (Sherbrooke, Canada).

### 4.2. Methodology outline

The methodology we use to acquire translations of medical terms is briefly outlined and schematized in Fig. 2. Starting from the acquisition of a corpus of parallel documents in two languages (here English and French) which we prepare as required, we align sentences, then words, and terms are finally extracted from the results and filtered. A list of terms and their translations is thus produced.

### 4.3. Parallel corpus: acquisition and preparation

Today a powerful resource for collecting a corpus is the Web. Besides providing access to an unlimited number of documents, it also hosts a large quantity of multilingual texts [11], which is our concern in this work. The only drawback to such a method is the difficulty to assess the quality of the documents, so that this might account for a proportion of noise in the results.

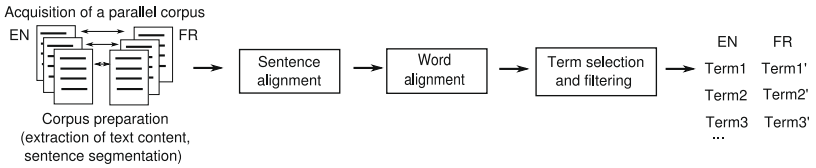


**Fig. 2.** Description of the methodology.

The issue is to locate and retrieve parallel web pages. It can be decomposed into three main problems: (1) how to identify a web site or web page as a starting point to collect the corpus, i.e. where to find translated web sites or web pages; (2) how to identify pairs of parallel pages, i.e. pair each page in a language with its translation in the other(s) language(s); (3) how to check the quality of the obtained page pairs (i.e. are they indeed parallel?). For each of these three sub-problems, methods can rely on several types of criteria that are potential indicators of parallelism [11]: (a) meta-information (file names and links between documents); (b) language (parallel documents must be in different languages); (c) content (similarity of content may indicate that two pages are parallel). We built our parallel corpus from the Web using a combination of those three types of criteria. The identification of a web site (step 1) was direct since we knew of one: the Canadian health website "Health Canada",[1] which is bilingual (English/French) and entirely parallel—each English document having its French translation. The pairing of parallel documents (step 2), i.e. linking an English document to its corresponding French translation, was done using information contained in the pages. Indeed each page had a hyperlink to its translation page (e.g. *<a* ***href="../../english/diseases/hepatitis.html"****>...</a>*). To filter potential errors and to ensure that documents were rightly listed as English or French (step 3), we also checked the language of the document using two sources: the meta-information in the documents (language is given in the meta tags, e.g. *<meta name="dc.language"* ***content="eng"****>*), and the URLs (some are labelled as "*english*" or "*french*", e.g. "www.hc-sc.gc.ca/english/diseases/hepatitis.html"). After downloading the whole website (version of January 2005), which resulted in 20,000 documents (10,000 pairs of parallel documents) and over 27 million words, we randomly selected, for reasons of processing time, a set of 760 documents pairs and 1.29 million words which provided the basis for this experiment. Thus we obtained a medical corpus of parallel English and French documents to be processed with alignment tools.

Web pages are generally HTML documents where text content is mixed with and surrounded by markup which is of limited use to Natural Language Processing. Therefore, the textual content of the documents needs to be extracted. Several problems make this extraction a non-trivial task. HTML tagging may be erroneous (incorrect HTML) or incomplete (as allowed in the standard). Different character encodings may be used in different documents. Which parts of the document are actual text content (rather than, e.g. Javascript programs) must be decided. To convert our set of HTML documents retrieved from the web to text, we opted for two steps dedicated to two conceptually different tasks. First, cleaning of the HTML markup and conversion to a normalized form using the XHTML standard was done with the Web page cleaner HTML Tidy.[2] Second, selection of text content was performed on this normalized form, using an XSLT stylesheet we designed. This method provides easy access to the content and structure of the documents. While converting the corpus to text format, we decided to keep a number of tags that we thought would help during the next step (i.e. sentence alignment) as they are likely to be points of correspondence. Indeed a title in English is likely to correspond to a title in French, a hyperlink to a hyperlink and in most cases a paragraph to a paragraph. Thus title, paragraph and link tags were selected and extracted together with the text content to form the (lightly tagged) text version of the corpus. Finally, the corpus was also segmented into sentences as a necessary preliminary step to alignment.

### 4.4. Sentence alignment

As the complexity of word alignment in two strings of words tends to increase with the length of the strings, it is wise and common practice to reduce this length by first aligning texts at the level of sentences (i.e. pairing together sentences that are translations of each other). The issue to solve in sentence alignment is that there is generally not a one-to-one correspondence between all sentences of two parallel texts, as emphasized in Section 2. Sentence alignment techniques may rely on statistics or on linguistic knowledge. When statistical [12], they are often based on sentence length, assuming that parallel sentences have related lengths, which are measured either in characters or in words. When linguistic [13], methods usually look for similar words in sentences, using bilingual lexicons and searching for words with spelling similarities (referred to as "cognates"). State-of-the-art approaches often combine linguistic and statistical techniques [14]. To align the sentences of our corpus, we used a state-of-the-art statistical and linguistic aligner—GMA[3] (Geometrical Mapping and Alignment) [15]. We provided GMA with a French–English bilingual lexicon.

Though sentence aligners usually are robust tools that achieve high performances, any mistake at this level will be reflected at the next one—i.e. word alignment. The same holds for the pairing of parallel documents which was performed during corpus acquisition: we know the corpus to be parallel and expect the pairing of documents to be generally correct, but a small proportion of wrong pairs may have been included. So, in order to work on cleaner data, we designed a method to filter out possible noise by assessing the quality of sentence alignment: (1) we tried to detect incorrect sentence alignments by comparing sentence lengths and removing sentence pairs with too different lengths; (2) we looked for bad document pairs (documents that are not parallel) by evaluating the quality of sentence alignment between two documents. Sentence alignment is indeed likely to achieve low performance on non-parallel documents. We gave a score to each pair of documents and discarded documents under a certain score. The score of the documents is calculated according to scores given to each different type of sentence alignments. As seen in Section 2 there can be different types of sentence alignments: one sentence can be aligned with one sentence, or with two sentences or more, or can even be omitted (the common cases being 1:1 alignments). We gave a score to each alignment type, penalizing the most unlikely alignments (omissions and many-to-many alignments). The final score reflecting the alignment of the document should be close to 1 and we empirically set the threshold for selection to 0.66.

[1] www.hc-sc.ca.
[2] http://tidy.sourceforge.net.
[3] http://nlp.cs.nyu.edu/GMA/.

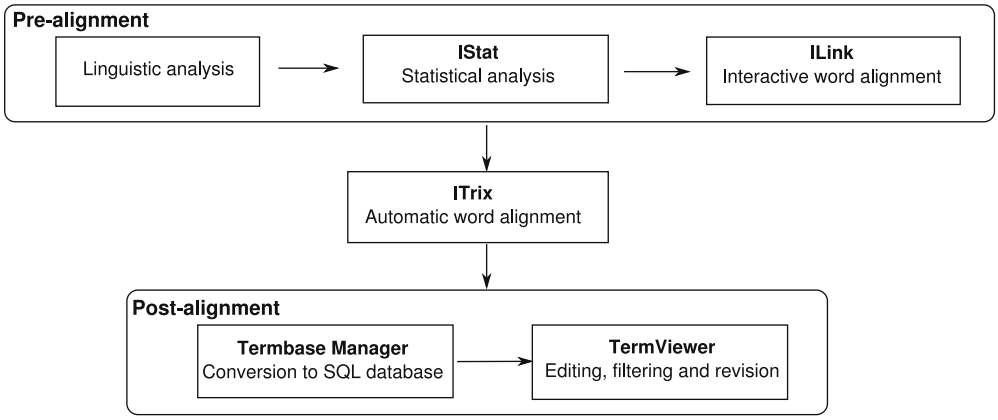


**Fig. 3.** Alignment with the ITools.

### 4.5. Word alignment

As explained in Section 2, word alignment is hard because sentences are generally not translated word by word. An additional difficulty is that morphological and grammatical properties may vary in different languages: word order, use of determiners, case marking, to cite but a few, may be quite different from one language to another.

In the same line as sentence alignment systems, word alignment systems can be based on statistical or linguistic techniques, or a combination of both. Statistically based methods rely either on association measures [16], i.e. cooccurrence measures: they look for pairs of words frequently occurring in corresponding sentences, the hypothesis being that words in translation relation occur in corresponding sentences more frequently than other words; or they build probabilistic translation models [17] (i.e. they compute more precisely the probability that a given word in a sentence be a translation of another word in the aligned sentence). They are especially effective on large corpora with high-frequency words but performances decrease with low-frequency occurrences. Linguistically based methods [18] make use of information such as syntactic parsing, bilingual lexicon and cognates. They are less robust despite being able to deal with low-frequency words. Hybrid approaches [19] seem to be a good compromise.

To align the words of our corpus, we used state-of-the-art tools, the Itools suite, which relies on both statistical and linguistic knowledge, trying to make the most of all kind of elements potentially helpful (statistical measures, POS tagging, syntactic parsing, etc.). This suite is composed of several modules: (1) **IStat**: a component calculating statistics on the corpus which serve as resources for the next components; (2) **ILink**: an interactive aligner used to train the automatic aligner; (3) **ITrix**: an automatic word aligner which constitutes the core of the Itools suite; (4) **Termbase Manager**: a component transforming the results of ITrix into an SQL database; (5) **TermViewer**: a graphical tool allowing the user to review, filter and categorise the results. These steps are schematized in Fig. 3 and detailed below.

#### 4.5.1. Pre-alignment

As stated above, the alignment tools exploit various information which are likely to help achieve better results. Automatic alignment may indeed not perform optimally without more resources. This leads to collect of linguistic and statistical data as a pre-alignment step. Training data learned on the corpus is also useful and is acquired through interactive alignment as another pre-alignment step. All the data thus gathered will serve as input data to the subsequent core task of automatically aligning the corpus.

Linguistic processing is performed on the corpus, as requested by the word alignment tools. Both the English and French halves of the corpus are thus tagged and lemmatized using the POS tagger Treetagger[4] [20] and syntactically parsed with Syntex [21], a dependency syntactic analyzer. The result of this linguistic analysis consists of two files containing annotated text in XML format as required by the tools.

The statistical tool IStat is applied to the two parallel XML files. This creates bilingual lexical resources based on statistical measures—cooccurrence measures. That is, the tool looks for pairs of words that are frequently found in corresponding sentences (i.e. sentences aligned during the sentence alignment step) and thus likely to be translations of each other.

Like IStat, the interactive word aligner ILink [22] is another component that is used to create resources for the automatic aligner. It allows to learn training data from the corpus, thus helping the automatic aligner perform better. Bilingual resources are built incrementally each time words (or terms) are aligned by the user. ILink is a graphical interactive tool (see Fig. 4) that enables the user to control the alignment process. Alignments (in corresponding colors) are proposed to the user who can accept, reject or revise them. Both positive ("Accept" decision) and negative ("Reject" decision) data are stored. The software is able to "learn" from the decisions of the user and performances improve as the training goes on.

Aside from the data created with IStat and ILink, other resources such as a bilingual lexicon and parts-of-speech correspondences (e.g. *Adjective + Noun* in English corresponds to *Noun + Adjective* in French) can also be used.

#### 4.5.2. Automatic word alignment

Once pre-alignment steps are completed, automatic alignment, which constitutes the core issue of word alignment, is performed. All the resources compiled at the previous stage (statistical resources, data from ILink, bilingual lexicons...) are used together with the automatic alignment component, ITrix. This module is fully automatic and does not require any intervention from the user, apart from configuring the files and how the resources are going to be applied.

#### 4.5.3. Post-alignment: conversion, filtering and revision

When the corpus has been aligned, individual alignment results are scattered across the whole corpus. Moreover all alignments are not equally reliable and human intervention is necessary to review and validate the results. We need to gather similar alignments

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

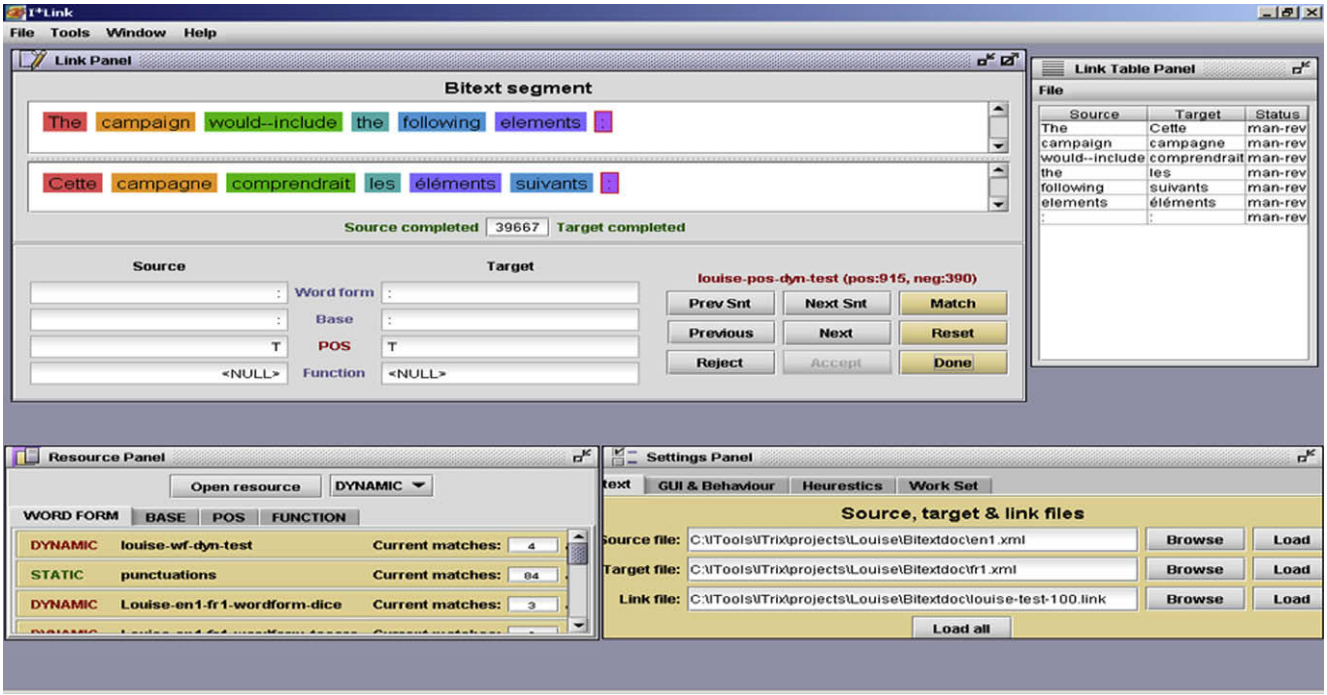


**Fig. 4.** Interactive word aligner.

together, order them, assess their quality to filter out potential bad results and present the results to the user. To do so, the output is converted into a relational database thanks to the Termbase Manager. It gathers together similar alignments so as to output only type alignments—that is to say inflectional variants of the same word are brought together under the same type, e.g. the words *patient* and *patients* corresponds to the same type. The tool also stores information such as inflectional variants (corresponding to the types), parts-of-speech and examples from the corpus.

After the conversion to a term database, the TermViewer interactive tool (see Fig. 5) can be used to filter, revise and categorise the terms. The user is presented with the set of alignments (i.e. source words paired with their target translations) along with their frequencies, parts-of-speech, status, quality scores, inflectional variants, as well as example sentences to present the words in their context. The user can revise each pair of words, accepting or rejecting the alignment. The alignments can be sorted according to several criteria such as alphabetical order, status, frequency and quality score. The quality score (referred to as *Q*-value) is a ranking value designed to predict the quality of alignments and is defined as follows:

$$Q = \frac{f_{st}}{(n_s + n_{t)}}$$

where $f_{st}$ is the frequency of a given type pair (a source word type $s$ and a target word type $t$), and $n_s$ and $n_t$ the number of different type pairs (obtained after alignment) in which the source and target types occur, respectively.

This measure, close to the Dice coefficient association measure [23], compares the frequency of the types as a pair with the frequencies of each type of the pair independently. The underlying hypothesis is that word pairs with high frequencies and consistent translations are of higher quality than pairs with low frequencies and exhibiting great variations on either the source or the target side. The *Q*-value is a good indicator of the accuracy of alignments [24] and selecting only the alignments above a certain *Q*-value would therefore help filter out potential noise.

### 4.6. Term selection

The tools align the whole corpus, however our objective is terminological and we are only interested in terms belonging to the chosen medical terminologies, as opposed to work aiming at extracting unknown terms. We therefore need to detect them at some point. Two main approaches are possible when dealing with term alignment. Either (1) first extract the terms and second align them [25]; this approach is directed by its terminological objective and, in our case, will ensure to detect all terms in the English side of the corpus (since they belong to existing terminologies). However this will not be the case for the French side of the corpus where the terms are unknown and results will be dependent on the performance of a term extractor. Besides, if only the terms are kept, the alignment might be less good because of the loss of information. Or (2) first align the corpus and then extract the terms from the results [26]. This method keeps all the information contained in the texts and is likely to produce a good quality alignment. It is not dependent on language-dependent exterior tools to detect unknown terms. Yet since all the corpus is aligned, more processing is done than really necessary. Moreover since terms are not extracted beforehand, some occurrences of terms might be missed in the alignment process, thus reducing the coverage.

We opted for the second approach, aligning the whole corpus while trusting the alignment tools to detect multi-word units, and then selecting only the terms for which translations were wanted. Our choice was to favour the quality of alignment, and was possible because the alignment tools are able to align both single words and multi-word expressions. We then used the MetaMap program[5] [27] to match the English part of the alignment results to MeSH, Snomed CT and MedlinePlus terms. We obtained a list of such terms paired with their French translations.

Although the TermViewer component of the Itools suite allows filtering based on quality score and frequency results, we also

[5] We used the free downloadable version of MetaMap, MMTx version 2.4.C http://mmtx.nlm.nih.gov/.

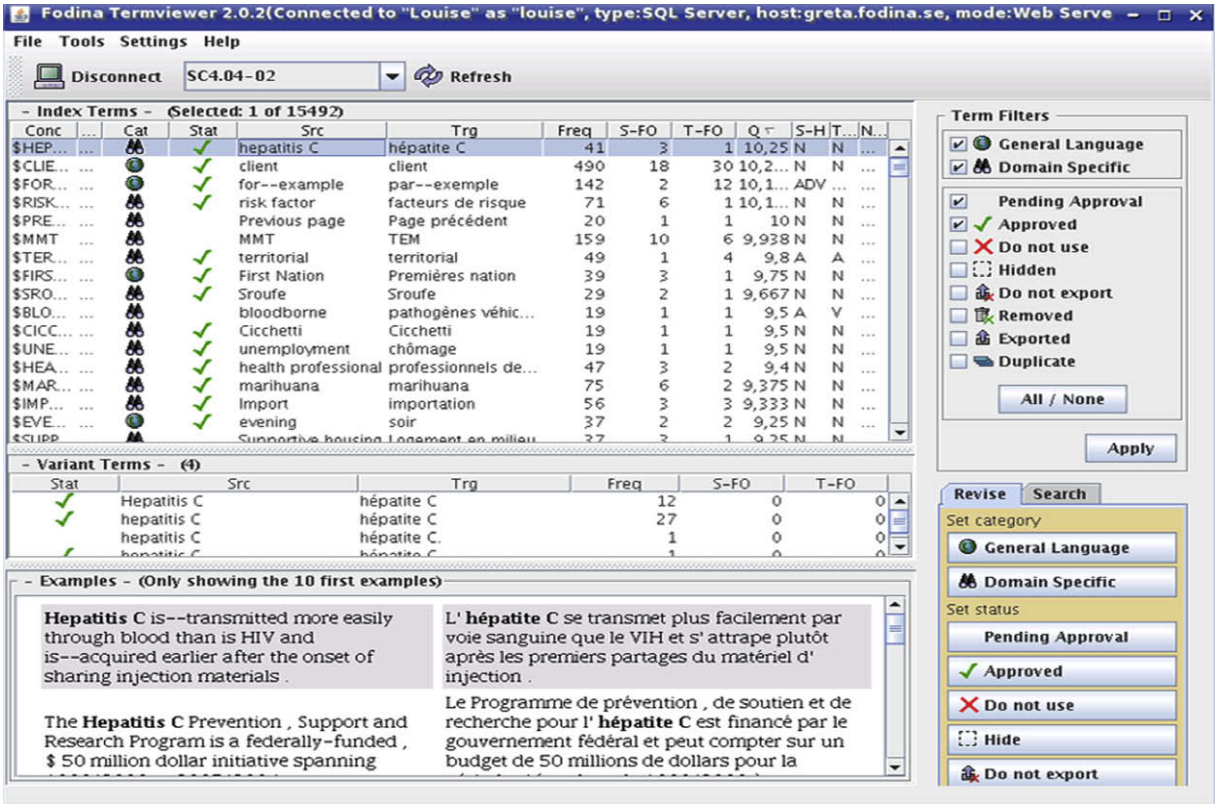


**Fig. 5.** Termviewer application—used for filtering and validation.

performed additional filtering customized according to our objective. Since medical terminologies mostly contain Noun Phrases and Adjectives, we chose to remove all conjugated verbs from the results, as we considered they might bring more noise than actually useful translations. Also, since part of the source terminologies are already translated, we can remove known translations from the results. Thus we removed: (1) translations that already existed in the French version of the terminologies; (2) since the three terminologies are part of the UMLS, we also removed French translations that could be obtained through the UMLS, i.e. French terms that were included in the UMLS and had the same concept identifiers (CUIs) as the English terms.

Once the terms are automatically selected and filtered, they are ready to be manually reviewed. An additional filtering step was also performed before doing so: we noticed that some words were not translated in French in the target texts, i.e. they were kept as is and thus gave English–English aligned word pairs in the results. While these types of alignments are not incorrect, they are nonetheless useless as we are looking for French translations of English words. We therefore removed them manually from the set of translations, as it would be difficult to remove them automatically since there is no sure indicator that a word has not been translated (perfect similarity between the words could be an indicator but since there are some English and French words that are spelled the same we cannot rely on this criterion).

### 4.7. Experimental setup

We described the general process of aligning terms to find translations of medical terms in Section 4.2, but there can be variations at the step of selecting terms according to the level of filtering that is performed. So we first experimented with different test setups detailed in previous work [28,29] which we sum up quickly. In a preliminary work [28] concentrating on MesH terms, we set up a simple experiment with no filtering aside from removing translations already. We then experimented with more advanced filtering (close to what is described here) [29], for instance we tested the use of the *Q*-value score to filter the results. This enabled us to set up a final implementation likely to give the best results. What differs also in this final implementation is the version of the ITools suite: we used a new version with additional features (conversion to database, gathering of similar alignments as type alignments). This rendered some of the filtering we tested in [29] obsolete.

The final implementation we designed and used in this work consists of the following steps:

(1) Sentence alignment (*automatic*)
(2) Word alignment
  (2.1) Linguistic and statistical analysis (*automatic*)
  (2.2) Training with ILink (*manual*): on 600 sentence pairs from the corpus. However training need not be repeated each time a corpus is processed. Here in particular, this step was skipped as we used the training resources of the test implementations.
  (2.3) Automatic alignment with ITrix (*automatic*)
  (2.4) Conversion to database (*automatic*)
(3) Term selection and filtering
  (3.1) Filtering with TermViewer (*automatic*): alignments with a *Q*-value superior to 0.4 were selected (this threshold was determined empirically after a few tests).
  (3.2) Detection of terms using Metamap (*automatic*)
  (3.3) Additional customized filtering (*automatic and manual*): we removed conjugated verbs, known translations and English–English word pairs.
(4) Review of the results (*manual*): the validation of the results is twofold. First, validity of the alignments, i.e. checking whether the aligner has correctly paired a term with its translation, is performed by a language engineer (and can potentially be performed by anyone with sufficient knowl-

edge of the two languages). Then medical accuracy of the translations has to be determined and this can only done by domain experts. By dividing the process of reviewing in two steps, we are able to reduce the burden on medical experts who only have to revise correct translations.

As can be seen, manual intervention is limited to a minimum in this setup.

### 4.8. Evaluation method

We evaluated the results according to two general criteria: quality and quantity, knowing that high performance in the former is the main focus of this work.

#### 4.8.1. Quality

The measure used for evaluating the quality of the alignment method was precision which is defined as: $P = (nbcorrectalignments)/(nbalignments)$

Evaluation was performed on three levels:

1. A general level: performance of the word aligner was evaluated on the overall alignment results as a way to investigate the performance of the ITools for aligning words. In order to exemplify the benefit of using the Q-value to filter the results and show that we can obtain good precision with this measure, we evaluated two samples: (i) **sample A**: 100 word pairs randomly selected from the unfiltered results (selected before step (3) of the experimental setup); (ii) **sample B**: 100 word pairs randomly selected from the results filtered with the Q-value (selected after step (3.1) of the setup). We reviewed the alignments ordered by Q-value (meaning that the alignments with the best Q-value were ranked and reviewed first). For each of the two sets of ordered alignments, we measured the overall precision, as well as precision at 10 different stages of the reviewing (precision for the first 10 alignments, then for the first 20 and so on).
2. The specific level of medical terms: evaluation as regards the linguistic validity of the alignments of medical terms was performed to examine the suitability of the method to the specific task of aligning medical terms. We evaluated the alignment for each of the three targeted terminologies. We reviewed the sets of selected and filtered terms (step (4) of the setup). We measured precision for each set of terms (MeSH, SNOMED and MedlinePlus) as well as for all terms together (removing duplicates).
3. The specific level of medical terms as regards the medical accuracy of the retrieved translations. Even if a medical term has been correctly aligned to its French translation in the corpus, this does not necessarily mean that this translation is desirable to be used in a medical terminology. The accuracy of the translation must also be assessed and this can only be done by medical experts working with terminologies. The translations we retrieved were too numerous to presented to experts, so we only submitted a sample of 283 translated MeSH terms for evaluation. We measured precision on this sample. It should be reminded that the accuracy of these results is dependent on the corpus rather than on the performance of the method.

#### 4.8.2. Quantity

We investigated the coverage of our method as regards the three source vocabularies and the terms from these vocabularies actually present in the corpus. We detected the English terms present in the corpus using MetaMap, as done for the detection of terms in the alignment results. We measured the number of different elements in the source terminologies and in the corpus at three different levels: the level of concepts (CUIs in the UMLS), the level of terms (LUIs in the UMLS), and the level of codes given in the source terminologies. We compared these figures to our results at two different steps of our method: after the selection of terms from the aligned pairs (i.e. the number of aligned terms at step (3.2) of the implementation) and after the reviewing of those pairs (i.e. the number of correct aligned terms). A special distinction is made for the MeSH thesaurus. This vocabulary actually contains a number of chemical names called the Supplementary Concept Records (SRCs) which are unlikely to be found in a text corpus such as ours. Therefore, if we are examining the coverage of this terminology it may be more meaningful to leave those elements out, which we did.

We also examined the coverage in terms of multi-word terms and single-word terms. We looked at the number of single-word and multi-word terms in the source terminologies and in the corpus, as opposed to their numbers in our results.

## 5. Results

Sentence alignment of the corpus resulted in 49,679 sentence pairs (step (1) of the experimental setup). 15,492 word alignments were obtained after filtering with TermViewer (step (3.1) of the setup). Among these 1042, 1566 and 131 alignments to be reviewed were extracted, respectively for the three terminologies MeSH, SNOMED CT and MedlinePlus (step (3.3) of the setup).

Evaluation of the overall performance of the word alignment tools (described in Section 4.8) gave a precision of 41% for sample A while sample B, with the filtered alignments (with a Q-value of 0.4 and above), obtained a precision of 79%. Fig. 6 displays preci-

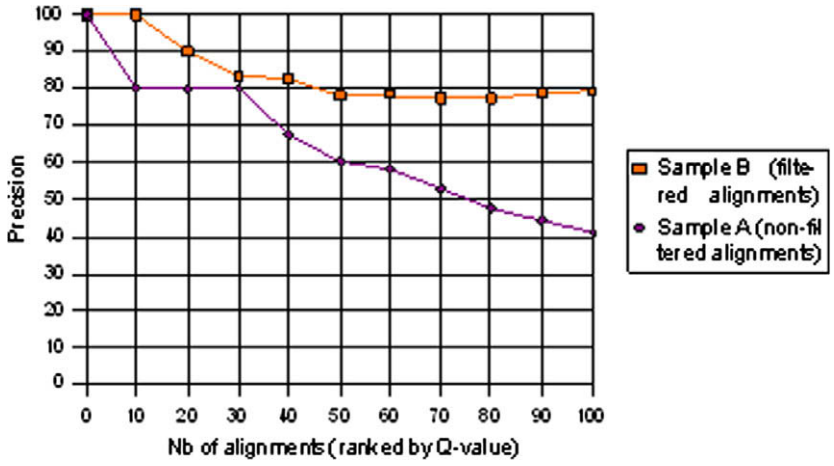


**Fig. 6.** Evolution of precision for samples A and B.

**Table 1**
Evaluation of alignment for the three medical terminologies.

| | MeSH | SNM CT | MedlinePlus | ALL |
|---|---|---|---|---|
| Alignments | 1042 | 1566 | 131 | 2127 |
| Correct alignments | 779 | 1218 | 100 | 1653 |
| Precision (%) | 74.8 | 77.8 | 76.3 | 77.7 |

**Table 2**
Examples of term translations.

| English | French | Source vocabulary |
|---|---|---|
| Breast milk | Lait maternel | MeSH |
| Sustainable development | Développement durable | MeSH |
| Phenolics | Phénoliques | SNOMED CT |
| Methadone maintenance | Entretien à la méthadone | SNOMED CT |
| Second-hand smoke | Fumée secondaire | MedlinePlus |
| Frostbite | Engelure | MedlinePlus |
| Reproduction rights | Droits de reproduction | MeSH |
| Blind | Store | MedlinePlus |

sion measured at 10 different points of the alignments ranked by *Q*-value for both samples. Precision for sample A is slightly decreasing for the first alignments but rapidly falls off from the 30th alignment to very low levels of precision. The point where precision starts to radically drop corresponds to alignments with a *Q*-value around 0.39. This highlights the fact that alignments with high *Q*-values have a high precision while those with low *Q*-values are in general bad alignments and cause the precision to drop. This advocates the need to filter low-ranked alignment and a *Q*-value of 0.4 seems to be a good threshold for selecting the best alignments. It can be clearly seen in the evolution of the precision for sample B: precision for the alignments with a high *Q*-value is indeed only slowly decreasing and remains (and even slightly increases) at a high level of precision (around 79%) from the 80th alignment. The fact that it stops decreasing towards the end is explained by the presence of low-ranked but correct alignments (with a *Q*-value under 0.4) that had previously been validated and so were not filtered out. The results of the evaluation of word alignment underline the benefit of using the *Q*-value score to filter and rank the alignments, and show that with this filter the performance of word alignment is very good.

A proportion of the noise in word alignment can be attributed to errors in the sentence alignment process. Other factors not due to the word aligner include errors in POS tagging, bad document pairing and low quality of the data (misspelling of words, missing spaces between words...). Errors made by the alignment tools are of two types: (1) partial errors: a part of the alignment is correct but not all; (2) full errors: none of the aligned words corresponds to the other.

Table 1 shows the precision for the sets of terms from each terminology, as well as for all medical terms gathered together (removing duplicate terms). Although known translations have been filtered out, thus reducing the proportion of correct results, we see that figures are quite good for word alignment results. Examples of translations are given in Table 2. The last two lines show non-valid examples: they are linguistically correct but they correspond to different meanings of the English source words. 174 terms from the sample of MeSH translations submitted to experts (translators of the French MeSH) were validated as desirable additions to the French MeSH, which gives a precision of 61.5%.

The number of translations obviously depends on the size of the source vocabulary and on the number of terms actually present in the corpus. Table 3 displays the coverage of the source vocabularies and of terms present in the corpus. Each line indicates the number of different codes, concepts and terms in the vocabularies, in the corpus and in the alignment results (before and after reviewing). The ratio of one line over the previous one is also given, as well as the global coverage, that is, the number of correct alignments over the number of elements in the source vocabularies. We can see that the number of terms actually present in the corpus is low for the three terminologies (3194 for the MeSH, 4210 for SNOMED and 375 for MedlinePlus) and further decreases in the alignment results, especially in the final filtered and reviewed alignments. The small quantity is reflected through the global coverage ratio (respectively, 0.8%, 0.1% and 5.6% of terms). Terms from the MedlinePlus vocabulary are less numerous than for the MeSH and SNOMED but this vocabulary is very small (only 1387 terms) and the ratio of terms in the corpus and thus the global coverage are actually higher for this vocabulary than for the other two.

Table 4 shows the number of multi-word and single-words terms in the source vocabularies, in the terms present in the corpus and in the results (correct alignments). Except for the MedlinePlus Health Topics where the number of multi-word and single-word

**Table 3**
Coverage (number of different codes, concepts (CUIs) and terms (LUIs)).

| | MeSH (no SCRs) | | | SNOMED CT | | | MedlinePlus | | |
|---|---|---|---|---|---|---|---|---|---|
| | Codes | CUIs | LUIs | Codes | CUIs | LUIs | Codes | CUIs | LUIs |
| Source voc. | 23,997 | 45,622 | 79,848 | 368,593 | 299,907 | 788,204 | 708 | 1271 | 1387 |
| Corpus | 2676 | 3084 | 3194 | 4246 | 3980 | 4210 | 282 | 362 | 375 |
| % | 11.2 | 6.8 | 4 | 1.2 | 1.3 | 0.5 | 39.8 | 28.5 | 27 |
| Alignments | 1386 | 1526 | 1552 | 1927 | 1886 | 1975 | 170 | 196 | 200 |
| % | 51.2 | 49.5 | 48.6 | 45.4 | 47.4 | 46.9 | 60.3 | 54.1 | 53.3 |
| Correct align | 555 | 617 | 622 | 970 | 958 | 983 | 70 | 77 | 78 |
| % | 40 | 40.4 | 40.1 | 50.3 | 50.8 | 49.8 | 41.2 | 39.3 | 39 |
| Global coverage (%) | 2.3 | 1.4 | 0.8 | 0.3 | 0.3 | 0.1 | 9.9 | 6.1 | 5.6 |

**Table 4**
Number of multi-word (M-w) and single-word (S-w) terms.

| | MeSH (no SCRs) | | | SNOMED CT | | | MedlinePlus | | |
|---|---|---|---|---|---|---|---|---|---|
| | M-w | S-w | Total | M-w | S-w | Total | M-w | S-w | Total |
| Source voc. | 53,534 | 26,314 | 79,848 | 758,033 | 30,171 | 788,204 | 912 | 475 | 1387 |
| Corpus | 1154 | 2040 | 3194 | 1212 | 2998 | 4210 | 176 | 199 | 375 |
| Alignments | 212 | 410 | 622 | 249 | 734 | 983 | 40 | 38 | 78 |

terms is nearly similar, we see that in the resulting correct alignments the proportion of single-word terms is higher than that of multi-words terms, while it is the contrary in the source vocabularies.

## 6. Discussion

Alignment of the parallel corpus brought new translations of medical terms with a rather good quality. Indeed the precision rate for the alignment of these terms is high (a means of 77.7%) considering the complexity of word alignment. We should also stress the fact that the results are far better than with the implementations tested in earlier work [28,29]. Besides being linguistically correct translations, a fair portion of these translations (61.5% of our test sample) also seem to be desirable additions to a medical terminology.

Using the same alignment method as ours, Nyström et al. [5,6] aligned parallel terminologies to build a medical dictionary as opposed to text corpora. While this type of approach allows to process cleaner and medically appropriate data (since all words are part of medical terminologies), it also brings a lexical limit as only words already existing in terminologies will be aligned. The use of text corpora offers a more extensive field to look for translations, even if it means processing more unnecessary text. Also, we obtained similar levels of precision for the alignments as [5] even though we dealt with noisier data.

Moreover, the alignment performed here consisted of both single and multi-word units and the alignment tools were not limited to a particular type of multi-word units. While some methods concentrated on a specific term pattern (e.g. Adjective Noun [9]), we could detect various types of complex terms (e.g. Adjective Noun, Noun Noun, Noun Prep Noun...).

While the quality of translations is quite good, as shown by the precision figures, the quantity is rather low. There are several reasons for that. The size of our corpus is rather small (a total of only 1.2 million words). Should we process the whole Health Canada website (which consists of 27.7 million words), we would certainly acquire many more different translations. We cannot expect a direct proportion though (i.e., 23 times more different word pairs) because of the properties of text corpora (Zipf law, LNRE distributions [30]), but instead a gradual decrease in the percentage of new pairs.

The number of translations also depends on the type of corpus in relation with the type of source vocabularies. We used a health website aimed primarily at the general public which consequently brought more translations of lay vocabulary terms (MedlinePlus Health Topics). Specific types of corpora could be targeted according to the vocabularies for which translations are wanted. For instance, using article abstracts should be particularly suited to find translations of MeSH terms given the fact that the MeSH is a thesaurus for indexing scientific articles. In the case of our corpus, we have heterogeneous documents, with large sections dedicated to the general public but also some parts intended for medical specialists as well as some sections related to government health policies and legislations. Thus although the lay vocabulary obtained better coverage, the quantity of translated terms is nevertheless rather low and there is room for improvement. Characterizing and categorizing the content of the documents would allow us to select relevant sections to be processed depending on the target vocabularies and the proportion of terms should therefore rise.

Another reason involves multi-word terms. We relied on the automatic detection of multi-word units by the ITools to spot the multi-word terms considered as candidates to translation. This may miss occurrences of the multi-word terms that were present in our input vocabularies, thus missing potential translations. To avoid this, a different method would consist in first spotting all input terms in the English sentences, and either (1) take them into account in the alignment process (i.e. treat them as single units), or (2) keep the information in store in a list of detected terms and reassemble them after the alignment process (that is, their translations would be the sequence of French words paired with the sequence of English words that constitute the terms). This would ensure to detect all occurrences of English terms, however there are still some drawbacks. Nothing ensures that all occurrences will be aligned, some of them might be missed or dismissed because of low alignment quality. Treating multi-word terms as single units (method 1) might lower the quality of word alignment, especially if terms are detected only in the English side of the corpus (we ran preliminary tests with this method which performed poorly). The second approach would be especially challenged if some parts of the terms were not aligned, thus causing the reassembling to fail. Testing the second approach and comparing it with the approach presented in this article would provide more insights into the pros and cons of these two methods.

Finally, filtering the results also causes the quantity to drop, however since it is mostly incorrect translations that are removed, this step is desirable to obtain good quality results.

An advantage of this method is that it is automatic for the most part and saves time compared to a fully manual approach consisting in employing human translators. Although manual work is still required, it remains limited: training of the word aligner only needs to be done once and the task of reviewing the results is greatly alleviated thanks to the filtering process. Moreover the reviewing work is divided into two parts: validation of the alignments, which does not require domain expertise, and validation of the medical accuracy of the translations and of their suitability for inclusion in a thesaurus, which is the only step requiring experts.

Aside from saving time for a translator, this type of method also provides access to previously translated texts. Instead of starting from scratch, we can re-use previous work and identify attested translations that a human translator might not have thought of, especially if working to translate terminologies without textual context.

Finally, we should point out that this method may be applied directly to other language pairs than English–French, as long as parallel data is available. Indeed [5,6] used the same alignment tools as here but with English and Swedish. The method also provides a methodological framework that may be used with different processing tools, though the results may not be as good, depending on the performance of these tools.

## 7. Conclusion

To summarize, we were able to acquire new translations of English terms from three medical terminologies—the MeSH, SNOMED CT and the MedlinePlus Health Topics—by aligning the words of a parallel English–French text corpus. These translations obtained good precision rates from both a linguistic and a medical point of view. The method is also applicable to other languages. Prospects for this work include characterization of the text content to better target specific terminologies, as well as detection of medical terms before the alignment process and increasing the quantity of acquired translations.

## Disclosure

Author Magnus Merkel is co-owner of the company Fodina Language Technology AB, which holds the rights to commercial exploitation of the ITools suite.

## Acknowledgment

We thank Huguette Vallée (INSERM, DISC) for reviewing the candidate MeSH translations.

## References

[1] Lovis C, Baud R, Rassinoux AM, Michel PA, Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. Artif Intell Med 1998;14:201–14.
[2] Marko K, Baud R, Zweigenbaum P, et al. Cross-lingual alignment of medical lexicons. In: Proceedings of LREC 2006 workshop on acquiring and representing multilingual, specialized lexicons: the case of biomedicine. Genoa, Italy; 2006.
[3] Claveau V, Zweigenbaum P. Translating biomedical terms by inferring transducers. In: Silvia Miksch, Jim Hunter EK, editors. Proceedings of the 10th conference on artificial intelligence in medicine Europe, LNCS, vol. 3581. Berlin/Heidelberg: Springer; 2005.
[4] Baud RH, Lovis C, Rassinoux AM, Michel PA, Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. In: Cesnik B, Safran C, Degoulet P, editors. Proceedings of the 9th world congress on medical informatics; 1998. p. 581–5.
[5] Nyström M, Merkel M, Ahrenberg L, et al. Creating a medical English–Swedish dictionary using interactive word alignment. BMC Med Inform Decis Mak 2006;6:35.
[6] Nyström M, Merkel M, Peterson H, Ahlfeldt H. Creating a medical dictionary using word alignment: the influence of sources and resources. BMC Med Inform Decis Mak 2007;7:37.
[7] Chiao YC, Zweigenbaum P. Looking for French–English translations in comparable medical corpora. J Am Med Inform Assoc 2002;8(Suppl.):150–4.
[8] Widdows D, Dorrow B, Chan C. Using parallel corpora to enrich multilingual lexical resources. In: Proceedings of the LREC, Las Palmas, Spain: ELRA; 2002. p. 240–4.
[9] Ozdowska S, Névéol A, Thirion B. Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. In: Proceedings of the 6èmes rencontres TIA.
[10] Déjean H, Gaussier E, Renders JM, Sadat F. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. Artif Intell Med 2005;33(2):111–24.
[11] Resnik P, Smith N. The Web as a parallel corpus. Comput Linguist 2003;29:349–80. Special Issue on the Web as a Corpus.
[12] Gale WA, Church KW. A program for aligning sentences in bilingual corpora. Comput Linguist 1993;19(3):75–102.
[13] Kay M, Röscheisen M. Text-translation alignment. Technical report, Xerox Palo Alto Research Center; 1988.
[14] Simard M, Plamondon P. Bilingual sentence alignment: balancing robustness and accuracy. Mach Trans 1998;13(1):59–80.
[15] Melamed ID. Bitext maps and alignments via pattern recognition. In: Véronis J, editor. Parallel text processing: alignment and use of translation corpora. Dordrecht: Kluwer Academic Publishers; 2000.
[16] Fung P, Church K. K-vec: a new approach for aligning parallel texts. In: Proceedings of the 15th international conference on computational linguistics; 1994. p. 1096–102.
[17] Brown P, Pietra S, Pietra V, Mercer R. The mathematics of statistical machine translation: parameter estimation. Comput Linguist 1993;19(2):263–311.
[18] Wu D. Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammar. In: Véronis J, editor. Parallel text processing: alignment and use of translation corpora. Dordrecht: Kluwer Academic Publishers; 2000.
[19] Barbu AM. Simple linguistic methods for improving a word alignment algorithm. In: Proceedings of the 7th international conference on the statistical analysis of textual data JADT'04, Louvain-la-Neuve, Belgique; 2004. p. 88–98.
[20] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing. Manchester, UK; 1994. p. 44–9.
[21] Bourigault D, Fabre C, Frérot C, Jacques MP, Ozdowska S. Syntex, analyseur syntaxique de corpus. In: TALN, Dourdan, France; 2005.
[22] Merkel M, Petterstedt M, Ahrenberg L. Interactive word alignment for corpus linguistics. In: Proceedings of corpus linguistics 2003. Lancaster, UK; 2003. p. 533–42.
[23] Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press; 1999.
[24] Merkel M, Foo J. Terminology extraction and term ranking for standardizing term banks. In: Nivre J, Kaalep HJ, Muischnek K, Koit M, editors. Proceedings of the 16th nordic conference of computational linguistics NODALIDA-2007, University of Tartu, Tartu; 2007. p. 349–54.
[25] Gaussier E. Flow network models for word alignment and terminology extraction from bilingual corpora. In: Proceedings of the joint 17th international conference on computational linguistics and 36th annual meeting of the association for computational linguistics; 1998. p. 444–50.
[26] Foo J, Merkel M. Computer aided term bank creation and standardization. In: Steurs F, Thelen M, editors. Series terminology and lexicography research and practice. John Benjamins Publishing, Amsterdam, 2009. Under publication. Paper presented at the international conference on Terminology, Antwerp, 16-17 November, 2006.
[27] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. J Am Med Inform Assoc 2001;8(Suppl.):17–21.
[28] Deléger L, Merkel M, Zweigenbaum P. Enriching medical terminologies: an approach based on aligned corpora. Stud Health Technol Inform 2006;1:747–52.
[29] Deléger L, Merkel M, Zweigenbaum P. Contribution to terminology internationalization by word alignment in parallel corpora. In: Proceeding of the AMIA annual fall symposium, Washington DC; November 2006. p.185–9.
[30] Baayen H. Word frequency distributions. Dordrecht, Boston: Kluwer Academic Publishers; 2001.